Name: Sewade Ogun
Lab 1

Summary of Findings and Exploration on Detectron2.

1. Instance Segmentation

Model Architecture Used: I used the Mask R-CNN X101-FPN model (mask_rcnn_X_101_32x8d_FPN_3x). It is made up of ResNext-101 backbone architecture with FPN. It was trained on the Common Objects in Context (COCO) dataset with 1.5 million object instances and 80 object categories.

Examples of Segmented Images (Correct)

Fig 1: Me and my friends at a wedding. It was able to detect the instances of persons, a small wrist-watch (clock) and the handbags.
Fig 2: Two brothers playing football. It was able to detect the instance of person (even with irregular positioning) and the football.

Examples of Segmented Images (Incorrect)

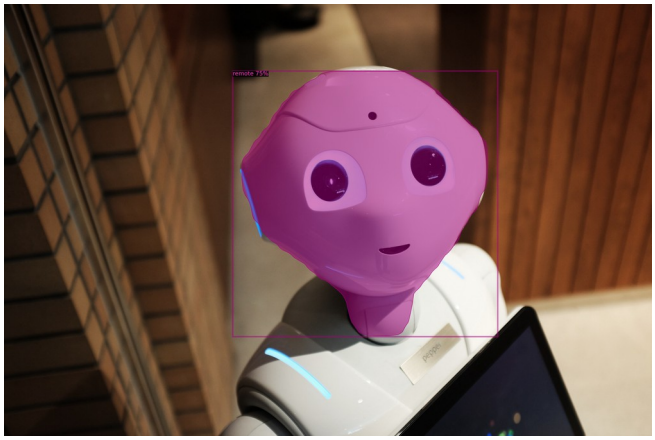
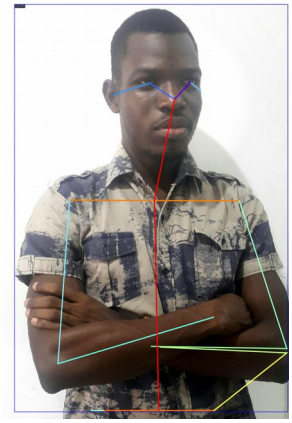


Fig 3: Female Players celebrating. It saw the red players' hand as hotdog.
Fig 4: A robot. The robot was labeled as a mouse.

2. Human Pose Estimation

Model Architecture Used: I used the Keypoint R-CNN X101-FPN model (`keypoint_rcnn_X_101_32x8d_FPN_3x`). It is made up of the ResNext 101 backbone architecture with FPN. It was trained on the Common Objects in Context (COCO) dataset with 250,000 people with keypoints.

Examples of Pose Estimated Images (Correct)



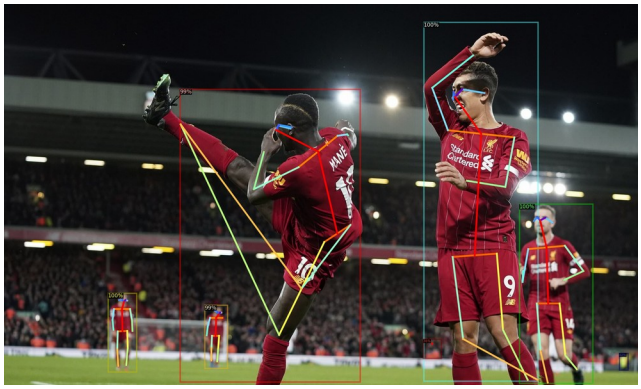


Fig 5: My portrait: A standing pose with hand-folded was predicted correctly
Fig 6: Sadio Mane (a footballer) celebrating a goal. The player's poses were predicted correctly, even with a leg raised.

Examples of Pose Estimated Images (Incorrect)

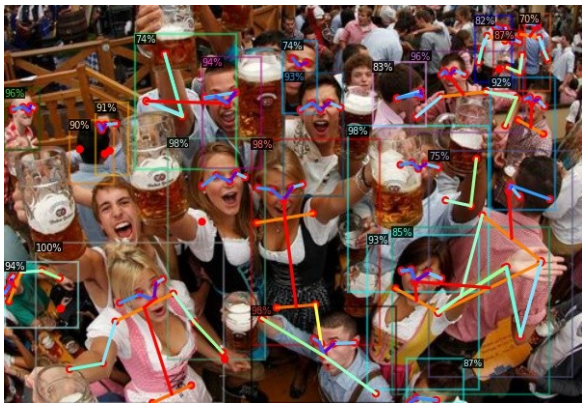


Fig 7: Children playing. The model could not capture the hand in the water due to the irregular position.
Fig 8: Crowd celebrating. Some poses were missed due to crowd and occlusion in the image while raising hands.

Observations
Instance Segmentation
The model can get tricked by occlusion and lighting. For example, in Fig 3 where the hand was seen as hot-dog. Also, the model fails in crowded situations where it is difficult to detect every person or some parts may be partially occluded as in Fig. 8.

Error modes
Error in predicting accurate poses for to-down view images. And also for images with crowd. The model can easily get confused with so many persons packed into a small space as it has to distinguish each person for segmentation and pose estimation. The model needs to be able to accurately segment crowds and deal with occlusions in body parts like hands, feet etc.